CLAIMS

What is claimed is:

1. A method for extracting data from a Web page comprising the computer-implemented steps of:

using natural language processing, finding possible formal names on a given Web page, the step of finding producing a first found set of formal names;

searching the given Web page for formal names not found by the natural language processing step of finding, said searching producing a second set of formal names; and

refining a combined set of formal names formed of the first found set and the second set, said refining producing a working set of people and organization names extracted from the given Web page.

2. A method as claimed in Claim 1 wherein the step of refining includes rejecting predefined formal names as not being people names of interest.

3. A method as claimed in Claim 1 wherein the step of refining includes determining aliases of respective people and organization names in the combined set, so as to reduce effective duplicate names.

4. A method as claimed in Claim 1 wherein the step of finding further finds professional titles and determines organization for which a person named on the given Web page holds that title.

5. A method as claimed in Claim 4 wherein the step of finding includes employing rules to extract at least title and formal names.

6. A method as claimed in Claim 1 wherein the step of finding further includes determining educational background of a person named on the given Web page, the educational background including at least one of name of institution, degree earned from the institution and date of graduation from the institution.

7. A method as claimed in Claim 1 wherein the step of finding further includes determining biographical information relating to a person named on the given Web page.

8. A method as claimed in Claim 7 wherein the step of determining biographical information includes determining current and previous employment history of the named person.

9. A method as claimed in Claim 1 further comprising the steps of:

determining type of the given Web page; and

from the determined type, defining contents of different portions of the Web page, such that the steps of finding and searching are performed as a function of the defined contents.

10. A method as claimed in Claim 9 wherein the step of determining type of the given Web page includes determining structure or arrangement of contents of the Web page.

11. A method as claimed in Claim 10 further comprising the step of using the determined type, deducing additional information regarding a named person or organization on the given Web page, the additional information supplementing information found on another Web page of a same Web site as the given Web page.

12. A method as claimed in Claim 1 wherein the step of finding further includes determining at least one of addresses, telephone number, and email address relating to a person or organization named on the given Web page.

13. A method as claimed in Claim 1 wherein the step of searching employs pattern matching.

14. A database having records formed by data extracted from Web pages by the method of Claim 1.

15. A method for extracting information from a Web page document comprising the computer implemented steps of:

performing a lexical analysis on a given Web page document to identify elements of interest, the elements of interest producing formal names;

detecting a regular recurrence of a certain type of element, the detecting producing additional formal names;

resolving aliases of the produced formal names and additional formal names to form a working set of names of people and/or organizations named in the given Web page document.

16. A method as claimed in Claim 15, further comprising the step of transforming the given Web page document into a standardized form, the step of transforming including identifying page structure of the Web page document.

17. A method as claimed in Claim 15, further comprising the step of assigning a type to each line in the given Web page document, the step of assigning a type indicating purpose of each line in the given Web page document.

18. A method as claimed in Claim 17 wherein the step of performing a lexical analysis further identifies elements of interest on lines of certain assigned types.

19. A method as claimed in Claim 17 wherein the step of detecting includes using pattern matching, detecting a regular recurrence of a certain type of line, to produce additional formal names.

20. A method as claimed in Claim 15 wherein the step of performing a lexical analysis includes syntactically and grammatically identifying elements of interest.

21. A method as claimed in Claim 20 wherein the step of identifying elements of interest identifies noun phrases that correspond to a person or organization named in the given Web page document.

22. A method as claimed in Claim 20 wherein the step of performing a lexical analysis includes using natural language processing.

23. A method as claimed in Claim 20 wherein the step of performing a lexical analysis includes utilizing rules describing composition of a name.

24. A method as claimed in Claim 15 wherein the step of resolving aliases includes employing rules for determining variant versions of a person's name or an organization's name.

25. A method as claimed in Claim 15 wherein the step of aliasing includes rejecting names containing predefined forms of common known phrases.

26. A method as claimed in Claim 15 further comprising the steps of:

grouping subsets of lines together to form respective text units; and

extracting from the formed text units desired information relating to the people or organizations named in the given Web page document

wherein the step of grouping identifies boundaries where information about a person or organization is to be found.

27. A method as claimed in Claim 26 wherein the step of grouping recognizes elements of information that span across more than one line.

28. A method as claimed in Claim 26 wherein the step of extracting includes:

determining type of Web page document; and

from the determined type, defining contents of different portions of the Web page document such that extraction is performed as a function of the defined contents.

29. A method as claimed in Claim 28 wherein the step of determining type of Web page document includes determining structure and organization of contents of the document.

30. A method as claimed in Claim 28 wherein the step of extracting includes determining whether the given Web page document is a press release, and if so, identifying organization mentioned in the press release.

31. A method as claimed in Claim 26 wherein the step of extracting includes using a parser to recognize the relationship between elements of information.

32. A method as claimed in Claim 31 wherein the step of extracting further includes utilizing predefined semantic frames for determining (i) sentences that express a

relationship between a person and organization named in the given Web page document and (ii) sentences that express that a person has a certain level of education.

33. A method as claimed in Claim 26 wherein the step of extracting includes associating a person or organization with an element of information if said element appears in a non-sentence within a formed text unit for that person or organization.

34. A method as claimed in Claim 26 wherein the step of extracting further divides a line that contains multiple names.

35. A method as claimed in Claim 26 wherein the step of extracting is rules based.

36. A method as claimed in Claim 15 further comprising the step of post-processing to extract further names of organizations and relationships to people named in the given Web page document.

37. A method as claimed in Claim 36 wherein the step of post-processing includes:

extracting organization names from professional titles held by a named person;

associating a named person with an organization whose Web site is hosting the given Web page document; and

deducing organization names from biographical text of a named person.

38. Computer apparatus for extracting information from a Web page comprising:

a source of Web pages of interest;

an extractor coupled to receive Web pages from the source, the extractor being computer implemented and using natural language processing to extract desired information from the Web pages; and

a storage subsystem coupled to the extractor for storing the extracted desired information in a data store.

39. Computer apparatus as claimed in Claim 38 wherein the extractor extracts desired information from a given Web page by:

using natural language processing, finding possible formal names on a given Web page, the step of finding producing a first found set of formal names;

using pattern matching, searching the given Web page for formal names not found by the natural language processing step of finding, said searching producing a second set of formal names; and

refining a combined set of formal names formed of the first found set and the second set, said refining producing a working set of people and organization names extracted from the given Web page.

40. Computer apparatus as claimed in Claim 39 wherein the extractor further determines aliases of respective people and organization names in the combined set so as to reduce effectively duplicate names.

41. Computer apparatus as claimed in Claim 39 wherein the extractor further finds professional titles and determines organization for which a person named on the given Web page holds that title.

42. Computer apparatus as claimed in Claim 39 wherein the extractor further determines educational background of a person including at least one of name of institution, degree earned from the institution and date of graduation from the institution.

43. Computer apparatus as claimed in Claim 39 wherein the extractor further determines employment history of a person named on the given Web page.

44. Computer apparatus as claimed in Claim 38 wherein the extractor is rules based.

45. Computer apparatus as claimed in Claim 38 wherein the extractor further determines type of the given Web page, and from the determined type defines contents of different portions of the Web page, such that extraction of desired information is performed as a function of the defined contents.

46. Computer apparatus as claimed in Claim 45 wherein the extractor further using the determined type, deduces additional information regarding a named person on the given Web page, the additional information supplementing information found on another Web page of the same Web site as the given Web page.

47. Computer apparatus as claimed in Claim 38 wherein the extracted desired information includes names of people or organizations named on the given Web page, addresses, telephone numbers and email addresses relating to the named person or organization.

48. Computer apparatus as claimed in Claim 38 wherein the storage subsystem is formed of a loader responsive to the extracted desired information, the loader post-processing the extracted desired information to refine the extracted desired information for storage in the data store.